Supplementary File 2

Table-1: Example for contingency tables built under the three hypotheses tested by TFactS

transcription factor (foxo3)	user: present	user: absent	total
sign-less regulation hypothesis:			
catalogue: present	**11**	57	**68**
catalogue: absent	32	6301	6333
total	**43**	6358	**6401**
Pval Fisher's test			$0.00e^{+0}$
sign-sensitive activation hypothesis:			
catalogue: present	**8**	58	**66**
catalogue: absent	35	2328	2363
total	**43**	2386	**2429**
Pval Fisher's test			$1.40e^{-4}$
sign-sensitive inhibition hypothesis:			
catalogue: present	**3**	63	**66**
catalogue: absent	40	2323	2363
total	**43**	2386	**2429**
Pval Fisher's test			$2.15e^{-1}$

In this example, 43 regulated genes form Terragni et *al.* (34) were submitted to TFactS. Eleven of them belong to the foxo3 signature (defined by 68 regulations) in the sign-less catalogue ($6,401$ regulations in total). Among these 11 genes, 8 matched the regulation type ("up" and "down"), whereas 3 were of the opposite type in the sign-sensitive catalogue (containing 66 foxo3 signatures among $2,429$ regulations).

Table-2: TFactS comparison with other tools.

Ref.	Expected TFs	TFactS	Tfm-Explorer	Core_TF	Crsd	oPOSSUM	Gsea C3	Gsea TFactS
(35)	FOXO3	1^{st}	-	32^{nd}	-	-	-	-
	MYC	17^{th}	77^{th}	22^{nd}	3^{rd}	7^{th}	-	-
(36)	MYC	1^{st}	61^{st}	6^{th}	9^{th}	-	-	-
(37)	STAT1	1^{st}	-	-	-	2^{nd}	-	-
(38)	P53	1^{st}	94^{th}	-	3^{rd}	-	-	-
(39)	SREBP	3^{rd}	-	-	42^{nd}	N.A	-	1^{st}
(40)	NF-κB	1^{st}	86^{th}	22^{nd}	90^{th}	1^{st}	16^{th}	1^{st}
(41)	NF-κB	2^{nd}	48^{th}	1^{st}	1^{st}	1^{st}	-	1^{st}
(42)	SREBP	3^{rd}	-	55^{th}	20^{th}	N.A	-	-
(34)	FOXO3	2^{nd}	-	21^{st}	-	-	-	-
	NF-κB	1^{st}	-	-	59^{th}	11^{th}	-	-
(43)	SF-1	3^{rd}	-	43^{rd}	-	N.A	3^{rd}	-
(44)	EGR-1	3^{rd}	62^{nd}	12^{th}	26^{th}	-	-	-
(45)	SMAD	1^{st}	-	-	12^{th}	N.A	-	-
(46)	HIF-1	1^{st}	-	3^{rd}	N.A	-	-	-
(47)	SP-1	15^{th}	1^{st}	56^{th}	N.A	12^{th}	-	21^{st}
(48)	E2F	3^{rd}	-	7^{th}	1^{st}	6^{th}	-	-
(49)	AP-1	13^{th}	30^{th}	-	67^{th}	-	-	6^{th}
	Total hits	**18/18**	**8/18**	**12/18**	**12/16**	**7/14**	**2/18**	**5/18**

Numbers represent the rank of the expected TF for each study, based on p-value or FWER p-value (GSEA) among the top 100 list of predicted TFs. For GSEA the rank represented here is the best rank in both positive and negative output lists. GSEA was run using either "c3" signatures or TFactS sign-less catalogue. (-) symbol means: not found and (N.A): not applicable (CRSD does not support mouse data; oPOSSUM does not integrate profiles for SREBP, SMAD and SF-1). Only significant TFs having nominal p-value ≤ 0.05 are listed in this table. See Table 2 (manuscript) for references of the studies, which are listed in the same order.

Comparison methods

Tools that can detect TF binding sites over-representation in a set of co-expressed gene promoters are compared here with TFactS (sign-less catalogue). Results are sumarized in Table 2.

The GseaPreRanked tool of GSEA (v2 GUI version) was run using the motif gene sets: "c3.all.v2.5.symbols.gmt" or TFactS sign-less catalogue. The parameters used with GSEA are as follows:
xtools.gsea.GseaPreranked, with plot top x = 10, norm = meandiv, scoring scheme = weighted, make sets = true, mode = Max probe, gmx = c3.all.v2.5.symbols.gmt, set min = 2, nperm = 1000, rnd seed = timestamp and set max = 500. Output ranking in GSEA is based on

FWER p-value. Outputs from GSEA are considered here as significant when having the nominal p-value ≤ 0.05. The minimum rank from either positive or negative output lists is considered in ranking comparison with other tools.

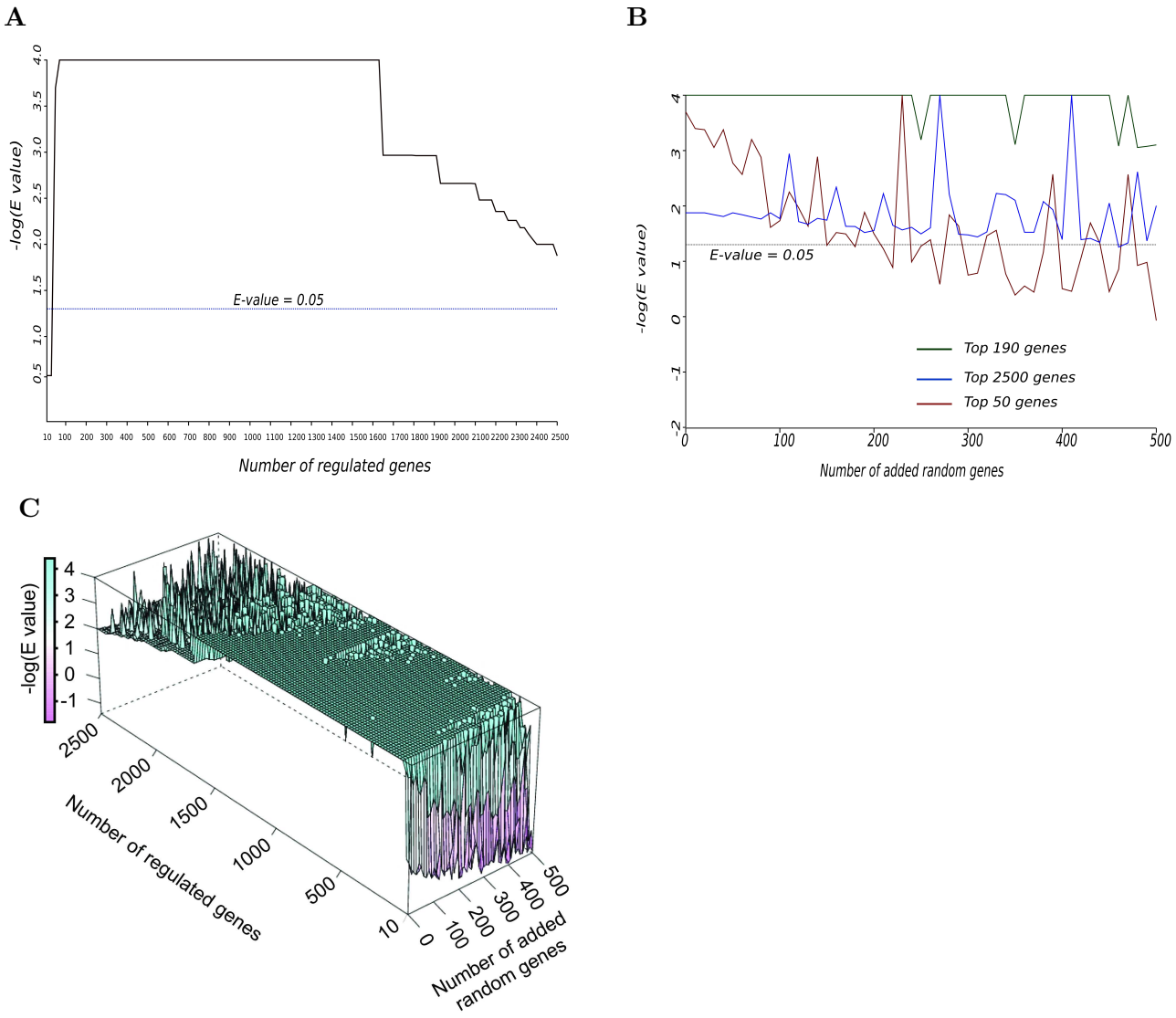
TFM-Explorer web server from `http://bioinfo.lifl.fr/TFME/` was run using both TRANSFAC and JASPAR vertebrate matrices on upstream promoter sequences for each gene located from -2000 to 0 with a ratio (density of clusters) of 2.5. Input gene sets were based on refSeq names. Significant TF binding sites were those with p-value ≤ 0.05. The minimum rank based on p-value is considered for a given TF in the comparison with other tools.

CORE_TF web server from `http://grenada.lumc.nl/HumaneGenetica/CORE\_TF` was fed with $2,000bp$ sequences upstream each gene first exon. The extraction of these sequences was done using Biomart web service (`http://www.biomart.org/biomart/martview/`). The random set generation was specified to be from Ensembl using 200 human gene promoters with the same base pair length upstream as the input promoters. The match cut-off selection was done with "Min False Sum" and significant TF binding sites were those with p-value ≤ 0.05 from the output list without processing for evolution conservation. The minimum rank based on p-value ranking is considered for a given TF when comparing with other tools.

oPOSSUM web server version 2 from `http://www.cisreg.ca/cgi-bin/oPOSSUM/opossum` was run by selecting the following inputs: Human as "Species", HUGO gene names for each set of co-expressed genes, vertebrates as "JASPAR core profiles", top 10% of conserved regions (min. conservation 70%) as the "Level of conservation", 80% as "Matrix match threshold" and 2000/0 as the "Amount of upstream / downstream sequence". Profiles with Fisher-Score ≤ 0.05 were considered as significant. The minimum rank based on Fisher-score is considered for a given TF for the comparison with other tools.

CRSD Genome-wide Iterative Enrichment Analysis web server from `http://140.120.213.10:8080/crsd/` was used with default values. And refSeq gene names were given as input. Significant TF binding sites were chosen based on 0.05 p-value threshold. The minimum rank based on p-value is considered for a given TF for the comparison with other tools.

For all these tools, the output lists that were chosen in this study were restricted to maximum 100 TFs if the generated significant lists contains more than 100 TFs.



Supplementary Figure 1: Robustness analysis of TFactS.

Significantly regulated genes in the EOL1 experiment were ranked according to the absolute value of the log2-ratio (untreated/treated). Then, the top 10 to 2,500 regulated genes were submitted to TFactS (sign-less).
Each run was repeated after addition of 10 to 500 genes randomly selected. The e-value score (-log10(e-value)) for STAT1 transcription factor is plotted against: **A**- The number of regulated genes. **B**- The number of randomly added genes to the top 50, 190 (genes regulated more than 3-fold) and 2500 regulated gene lists. **C**- The number of regulated genes and the overall numbers of randomly added genes.